THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 9, November 1983
Printed in U.S.A.

Performance of the Queueing Network Analyzer

By W. WHITT*

(Manuscript received March 11, 1983)

This paper describes the performance of the Queueing Network Analyzer (QNA), a software package developed at Bell Laboratories to calculate approximate congestion measures for networks of queues. QNA is compared with simulations and other approximations of several open networks of single-server queues. This paper illustrates how to apply QNA and indicates the quality that can be expected from the approximations. The examples here demonstrate the importance of the variability parameters used in QNA to describe non-Poisson arrival processes and nonexponential service-time distributions. For these examples, QNA performs much better than the standard Markovian algorithm, which does not use variability parameters. The accuracy of the QNA results (e.g., the expected delays) in these examples is satisfactory for engineering purposes.

I. INTRODUCTION AND SUMMARY

This paper is a sequel to Whitt,[1] which described the software package called the Queueing Network Analyzer (QNA). QNA calculates approximate congestion measures for networks of queues. The first version of QNA treats open networks of multiserver queues with the first-come, first-served discipline and no capacity constraints. QNA is designed to treat non-Markovian models: The arrival processes need not be Poisson and the service-time distributions need not be exponential. QNA approximately characterizes other kinds of variability through variability parameters assigned to each arrival process and each service-time distribution. The first step in the algorithm is

* Bell Laboratories.

to solve for the flow rates and the variahility parameters of tbe internal arrival processes. The second step is to compute approximate congestion measures for each queue separately by regarding it as a standard GI/G/m queue in which the renewal arrival process and the service-time distrihution are each partially characterized by tbeir first two moments or, equivalently, the rate and variability parameters. The third and final step is to calculate congestion measures for the network as a whole.

This paper describes the performance of QNA by comparing it witb simulations and other approximations of networks of queues. Even though QNA can analyze multiserver queues, only single-server queues are considered here. Among the other approximations in each case are the M/M/1 and M/G/1 approximations, which can he ohtained from QNA by using default options. The M/M/1 approximation, whicb is emhodied in the Markovian algorithms, is obtained by setting all variahility parameters equal to 1. With the M/M/1 approximation, the nodes are treated as independent M/M/1 queues with tbe correct rates. The M/M/1 approximation yields the exact equilibrium distribution of queue lengths for the Markov model with Poisson external arrival processes, exponential service-time distributions and one customer class. The M/G/1 approximation is ohtained by setting the variability parameter of each arrival process equal to 1 and using the specified service-time variahility parameter c_s^2; then the expected waiting time at each node is $(1 + c_s^2)/2$ times the M/M/1 value.

The congestion measures we consider in the examples here are the expected waiting time (hefore beginning service) and the expected sojourn time (waiting time plus service time) at a node or in tbe entire network. Of course, QNA produces other congestion measures, hut we are comparing with previously puhlished simulation results, which are mostly limited to expected waiting times and sojourn times.

We begin in Section II with a single GI/G/1 queue and discuss the implications of previous work on approximations for the GI/G/1 queue.[2–7] In Section III we consider a single queue with a superposition arrival process and compare QNA with simulations by Albin.[8–10] In Section IV we consider a network of eight queues in series analyzed hy Fraker,[11] and in Sections V and VI we consider two networks analyzed hy Kuehn:[12] Section V treats a tightly coupled two-node network and Section VI treats a nine-node network. In Section VII we treat a five-node network used to model a Bell Lahoratories computer system. Finally, in Section VIII we consider a model from Gelenbe and Mitrani[13] for a packet-switched communication network. The examples in Sections VII and VIII have input by classes and routes as in Section 2.3 of Whitt.[1]

These examples indicate the approximation quality tbat can be

expected in applications of QNA. They also demonstrate the importance of the variability parameters when the external arrival processes are not nearly Poisson or the service-time distributions are not nearly exponential. These examples also illustrate how to apply QNA, e.g., to model superposition arrival processes (Section III), to eliminate almost immediate feedback (Section V), and to conduct sensitivity analyses for the variability (Section VII).

II. A SINGLE GI/G/1 QUEUE

We begin by considering the special network containing a single service facility, in particular, the GI/G/1 queue with service times and interarrival times each partially characterized by their first two moments or, equivalently, by the four parameters τ (the mean service time), c_s^2 (the squared coefficient of variation of the service time), λ (the arrival rate), and c_a^2 (the squared coefficient of variation of the interarrival time).[1] The subscript indexing the node is suppressed since there is only one node.

It is useful to consider this model because it has been extensively studied and is relatively well understood. In many cases we can analytically determine the quality of the approximations for the GI/G/1 queue. Hence, we can get an idea about the quality of the approximations for more general networks. Of course, approximations for a node in a general network might be worse because the internal arrival processes usually are not actually renewal processes. On the other hand, the network operations of superposition and splitting tend to make stochastic point processes more like Poisson processes, so that larger networks may actually be better behaved.

The model specification above determines the approximate congestion measures produced by QNA, but the model specification is not complete since there are many service-time and interarrival-time distributions with the given parameters. The formulas produced by the QNA are approximations for all these systems, so it is natural to ask how the approximate congestion measures compare to the set of all possible values that are consistent witb tbe partial specification. Fortunately, it is often possible to identify the set of all possible values.[2–5] Moreover, it is often possible to locate the more likely values by identifying the set of all possible values under various natural constraints on the distribution. When this cannot be done exactly, it can often be done approximately using bounds.[6]

We now give a brief summary of results evaluating approximations for the expected waiting time or, equivalently (by Little's formula[14]), the expected queue length in the GI/G/1 queue based on tbe four parameters λ, c_a^2, τ, and c_s^2. First, recall that for the M/G/1 queue, with Poisson arrival process ($c_a^2 = 1$), the expected waiting time

actually depends on the service-time distribution only through the two parameters τ and c_s^2. Nonexponential interarrival-time distributions tend to be more difficult, however. For the GI/M/1 queue with an exponential service-time distribution ($c_s^2 = 1$), the expected waiting time depends on the interarrival-time distribution beyond the parameters λ and c_a^2. Given λ and c_a^2 in the GI/M/1 queue, the maximum relative error (upper bound minus lower bound divided by lower bound) in tbe mean queue length (number in system including anyone in service) is exactly c_a^2 (see Ref. 2). A similar, but somewhat less concise, result holds for the expected waiting time by virtue of Little's formula. The maximum relative error for the expected sojourn time (waiting time plus service time) is also c_a^2. This result suggests more generally that the reliability of the approximations migbt decrease when c_a^2 increases, which is consistent witb numerical experience.

If we assume that the interarrival-time distribution is not too irregular, then the maximum relative error becomes much less. In the $H_k/M/1$ queue with a hyperexponential interarrival-time distribution (mixture of exponential distributions having $c_a^2 > 1$), the maximum relative error in the mean queue length is $(c_a^2 - 1)/2$ (see Ref. 4). It turns out tbat the extremal interarrival-time distributions for $H_k/M/1$ queues also are extremal for all interarrival-time distributions that have Increasing Mean Residual Life (IMRL) (also $c_a^2 > 1$) and all service-time distributions,[5] so that the maximum relative error in the mean queue length for IMRL/G/1 queues is $(c_a^2 - 1)/[2 + \rho(c_s^2 - 1)]$.

Other kinds of shape constraints for the GI/M/1 queue have been investigated by means of nonlinear programming.[3] In general, we conclude that if the distributions are not irregular, then the maximum relative error in the GI/G/1 queue might be about $0.05\ c_a^2$, e.g., about 10 percent when $c_a^2 = 2.0$.

From heavy-traffic limit theorems that describe the queue as $\rho \to 1$,[6] where $\rho = \lambda\tau$ is the traffic intensity, we know that asymptotically the queue length and waiting-time distributions depend on the interarrival-time and service-time distributions only through the four parameters λ, c_a^2, τ, and c_s^2. This suggests tbat more generally the quality of the approximations might improve as ρ increases. This is certainly consistent with experience for the GI/G/1 queue, but not necessarily for more complex networks, e.g., the tightly coupled two-node network here in Section V.

The heavy-traffic limit tbeorems are closely related to diffusion approximations because diffusion processes emerge as limits in the heavy-traffic limit theorems. We have recently compared various diffusion approximations for tbe expected waiting time in a GI/G/1 queue to known bounds.[6] We now sbow how QNA and other related approximations fit into this framework. Table I here compares four

Table I—Bounds and approximations for the expected waiting time, EW, in a GI/G/1 queue: Three cases

	Parameter Values		
	$c_a^2 = 0.5$ $c_s^2 = 4.0$ $\rho = 0.7$	$c_a^2 = 2.0$ $c_s^2 = 4.0$ $\rho = 0.7$	$c_a^2 = 0.8$ $c_s^2 = 4.0$ $\rho = 0.3$
Daley's general upper bound	5.75	9.00	1.82
Monotone failure rate upper bound[5]	5.83	7.50	1.07
Kraemer and Langenbach-Belz[7]	5.14	6.88	1.02
QNA[1]	5.14	7.00	1.02
M/G/1	5.83 H	5.83 L	1.07 H
M/M/1	2.33 L	2.33 L	0.43 L
MFR lower bound[5]	5.00	5.83	0.93

Notes: 1. In each case the mean service time is $\tau = 1$.
2. "H" indicates high (greater than or equal to) and "L" indicates low (less than or equal to) in comparison with the bounds.

approximations for the expected waiting time, EW, with various upper and lower bounds in the three cases in Table 1 of Ref. 6. The four approximations are the M/M/1, M/G/1, Kraemer and Langenbach-Belz,[7] and QNA.

The M/M/1 approximation is obtained by replacing both variability parameters c_a^2 and c_s^2 by 1. The M/M/1 approximation is produced by a direct application of the Markovian software packages. The M/G/1 approximation is obtained by replacing c_a^2 by 1 and using the specified value of c_s^2. The M/G/1 approximation EW is the exact value for the approximating M/G/1 system since EW depends on the service-time distribution only through its first two moments. Both the M/M/1 and M/G/1 approximations are produced by QNA using default options.

The QNA approximation is the Kraemer and Langenbach-Belz approximation when $c_a^2 \leq 1$ and is slightly greater when $c_a^2 > 1$.[1] The one case in which $c_a^2 > 1$ in Table I shows that the difference between the two approximations is small compared to the distance between the upper and lower Monotone Failure Rate (MFR) bounds.[5] The MFR bounds are for interarrival-time distributions with decreasing failure rate when $c_a^2 \geq 1$ and increasing failure rate when $c_a^2 \leq 1$. The MFR bounds are tight when $c_a^2 \geq 1$ but not when $c_a^2 < 1$ (see Ref. 5). Since the interarrival-time distribution need not have monotone failure rate, the range of all possible values is greater, but the MFR bounds indicate the more likely values.

For the three cases in Table I, the M/M/1 approximation performs very poorly, falling way outside the bounds. The M/G/1 approximation always coincides with one of the MFR bounds—the upper bound when $c_a^2 \leq 1$ and the lower bound when $c_a^2 \geq 1$—but it would be better to have an approximation somewhere in the middle between the bounds. When $c_a^2 \geq 1$, the QNA approximation is a convex combination of the

two MFR bounds.[5] Since the MFR bounds are tight when $c_a^2 \geq 1$, the QNA approximation always yields the exact value of EW for some GI/G/1 system with the given parameters.[5] When $c_a^2 \leq 1$, the QNA approximation is slightly less than the convex combination of the MFR bounds. The convex combination of the MFR bounds is known to be an upper bound for $E_k/G/1$ systems,[5] so that it is appropriate to use a smaller value. We conjecture that the QNA approximation always yields an exact value of EW for some GI/G/1 system with the given parameters when $c_a^2 \leq 1$ too.

Table I provides a sample of the comparisons possible using the previous studies.[2-6] Since QNA coincides with the Kraemer and Langenbach-Belz approximation when $c_a^2 \leq 1$ and the Sakasegawa-Yu approximation when $c_a^2 \geq 1$, previous comparisons such as Tables 13 and 14 in Klincewicz and Whitt[3] also apply to QNA.

III. A QUEUE WITH A SUPERPOSITION ARRIVAL PROCESS

In this section we consider one single-server queue with a superposition arrival process. Such a system with two component arrival processes is depicted in Fig. 1a. Since only one external arrival process at each node is allowed in QNA, this model cannot be analyzed directly, but it is easy to modify the model so that QNA does apply. We added dummy nodes with very low traffic intensity on each component arrival process, as shown in Fig. 1b. Since the new dummy nodes have low

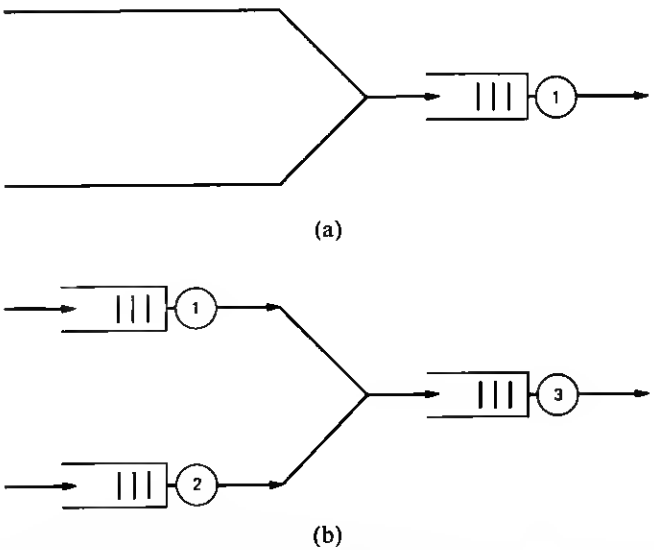

Fig. 1—(a) The original queue with a superposition arrival process. (b) The equivalent network with one external arrival process at each node.

traffic intensity, the rate and variability parameters of the departure processes from the dummy nodes will be almost identical to the corresponding parameters of the external arrival processes.

This model has recently been studied quite extensively by Albin[8-10] and is now relatively well understood. As in Section I, here we consider one illustrative example, which suggests the accuracy to expect more generally and shows the importance of the variability parameters.

The specific model we analyze is tbe $\Sigma GI_i/M/1$ system with an exponential service-time distribution and n iid stationary renewal processes as component arrival processes. The total arrival rate is 1, so the rate of each component process is n^{-1}. We consider six cases involving two values of n, $n = 2$ and 16, and three values of the traffic intensity ρ, $\rho = 0.3$, 0.7, and 0.9. In each case the component renewal-interval distribution is H_2^b, i.e., the mixture of two exponentials with balanced means: one with mean m_1 realized with probability p and the other with mean m_2 realized with probability $1 - p$, where $pm_1 = (1 - p)m_2$. In each case the squared coefficient of variation of the component renewal process interval is $c^2 = 6$. This is quite high variability, so that the component processes are not nearly Poisson. The three parameters of an H_2 distribution are determined by specifying the mean as n, $c^2 = 6$ and balanced means.

This model is taken from Chapter 3 and Appendix 6 of Albin.[8] Albin's approximations are based on the rate and variability parameters just as in QNA. In fact, the superposition approximation in QNA is a modification of Albin's procedure.[1] With Albin's procedure, the variability parameter of the superposition process is a convex combination of the variability parameters obtained from the stationary-interval method and asymptotic method described in Whitt.[15] The specific implementation of the stationary-interval method in Whitt[15] and Albin[8] is part of Kuehn's[12] algorithm for approximating networks of queues. Since Kuehn's implementation of the stationary-interval method is nonlinear, a different procedure is used in QNA. In QNA the variability parameter of the superposition process is a convex combination of the variability parameters obtained from the asymptotic method and a Poisson process.[1] Extensive experimentation has shown, however, that the approximation in QNA is very close to Albin's hybrid approximation, which performed very well in many experiments (about 3-percent average absolute relative percent error).

Several different approximations for the expected waiting time are compared with simulation in Table II. The simulation results were obtained in Albin.[8] The sample standard deviation is given in parentheses below the simulation estimate in Table II to indicate the statistical reliability of the estimate. The simulation program was written in FORTRAN using the "Super-Duper" subprogram in Mar-

Table II—A comparison of approximations and simulation of the expected waiting time, *EW*, in a ΣGI/M/1 queue with a superposition arrival process

No. of Renewal Processes, n	Traffic Intensity, ρ	Simulation Estimate	Approximation Method: M/M/1	Kuehn's Stationary-Interval Method	Asymptotic Method	Albin's Hybrid Procedura	QNA
2	0.3	0.205 (0.007)	0.128 (−37.6)	0.231 (+12.7)	0.281 (+37.1)	0.240 (+17.1)	0.236 (+15.1)
	0.7	4.57 (0.14)	1.63 (−64.3)	3.54 (−22.5)	5.32 (+16.4)	4.51 (−1.3)	4.64 (+1.5)
	0.9	26.3 (1.2)	8.1 (−69.2)	18.2 (−30.8)	28.2 (+7.2)	27.5 (+4.6)	27.5 (+4.6)
16	0.3	0.138 (0.004)	0.128 (−7.2)	0.147 (+6.5)	0.281 (+103.6)	0.153 (+10.9)	0.139 (+0.7)
	0.7	2.57 (0.07)	1.63 (−36.6)	1.88 (−26.8)	5.32 (+107.0)	2.36 (−8.2)	2.16 (−15.9)
	0.9	20.8 (0.94)	8.1 (−61.1)	9.4 (−54.8)	28.2 (+35.6)	21.1 (+1.4)	20.7 (−0.5)
Average Absolute Relative Percent Error			46.0	25.7	51.2	7.3	6.4

Notes: 1. The total arrival rate is 1 in each case.
2. The component renewal processes have H_2^b (hyperexponential) renewal-interval distributions with mean n, $c^2 = 6$, and balanced means.
3. The sample standard deviations appear below the simulation estimates in parentheses.
4. The relative percent error appears below the approximation values in parentheses.

saglia et al. to generate uniform random numbers.[16] A different random number seed was used for each simulation. The simulations began with an empty system, but the first 1000 customers were not counted to allow the system to approach steady-state. Each simulation consisted of 20 batches, with the number of customers per batch depending on the traffic intensity: 3,000 per batch for $\rho = 0.3$, 15,000 per batch for $\rho = 0.7$, and 50,000 per batch for $\rho = 0.9$. Even though much more simulation time was spent on the cases with higher traffic intensities, the statistical reliability was slightly less.

In Table II, in parentheses below the approximation values are the relative percent errors (RE), which are defined as

$$RE = 100(\text{Approx.} - \text{Simul.})/\text{Simul.} \tag{1}$$

At the bottom of Table II are the average absolute relative percent errors (ARE), which are defined as

$$ARE = \sum_{i=1}^{6} |RE_i|/6. \tag{2}$$

Dividing by the simulation value perhaps inflates the errors when $\rho =$

0.3 too much, but these summary measures provide a good overall comparison.

The stationary-interval method and the M/M/1 approximation are not bad for large n and small ρ because the superposition process converages to a Poisson process as $n \to \infty$ and the queue reflects this if ρ is not too big, in particular, if $n(1-\rho)^2$ is sufficiently large.[10,17] However, for $\rho = 0.9$, these two methods perform poorly. On the other hand, the asymptotic method performs reasonably well for $\rho = 0.9$, but not well in other cases. In particular, the asymptotic method does not reflect the convergence to a Poisson process as $n \to \infty$; it gives the same answers for $n = 2$ and 16. So, for fixed ρ, the asymptotic method gets worse as n increases.

As Albin determined in extensive experiments,[8,9] her hybrid approximation is much better than either basic metbod alone. This example also illustrates how close QNA is to Albin's hybrid procedure. For queues with superposition arrival processes, we conclude that QNA usually gives reasonable results and strongly dominates the two basic methods.

In closing this section, we add a caveat. The component processes in the simulation for Table II, in Albin's hybrid procedure and in QNA, are all based on the case of balanced means. However, as discussed in Whitt,[4,5] given the first two moments, the one-parameter family of renewal processes with hyperexponential renewal-interval distributions range from a Poisson process to a batch Poisson process with geometrically distributed batch size. For a single renewal arrival process, the expected waiting time, EW, in an $H_2/G/1$ queue also ranges between these same extremes, i.e., the M/G/1 and the $M^B/G/1$ systems. Unfortunately, no related theory yet exists for superposition arrival processes. However, we can easily describe what happens in the two extremes. The superposition of independent Poisson processes is Poisson and the superposition of independent batch Poisson processes with geometrically distributed batches having a common mean is batch Poisson with geometrically distributed batches. Thus, we *conjecture* that the maximum and minimum values for EW with a superposition of iid H_2-renewal processes correspond to the $M^B/G/1$ and M/G/1 systems, respectively. If the conjecture is true, then we would have the same range of possible values for H_2-superposition arrival processes as for H_2-renewal processes. However, if the component processes are not too batchy, then the superposition process will become more Poisson as n increases. We should usually expect the superposition process to be more nearly Poisson as n increases. To summarize, this heuristic analysis suggests that the range of possible values for EW in the $\Sigma GI_i/G/1$ queue given the basic parameters λ, c_B^2, τ, c_s^2 may be about the same as for the GI/G/1 queue. In fact, the

superposition operation may actually make the arrival process better behaved.

IV. EIGHT QUEUES IN SERIES

In this section we apply QNA to a network of eight single-server queues in series previously analyzed by Fraker.[11] The external arrival process is Poisson and all service-time distributions are Erlang. Fraker considered eight cases involving four traffic intensities ($\rho = 0.3$, 0.5, 0.7, and 0.9) and four Erlang service-time distributions ($M = E_1$, E_4, E_8, and $D = E_\infty$). Each of the traffic intensities and each of the service-time distributions are assigned randomly to two of the eight nodes. Fraker developed an approximation for these systems and compared it with simulations.

Tables III and IV describe Fraker's first two cases and the approximations for the expected waiting time at each node. The service-time squared coefficient of variation specifies the Erlang distribution since $c^2 = k^{-1}$ for E_k. Fraker made three simulation runs of 2500 customers, discarding the first 500 in each case to damp out the transient effects of starting the simulation. Statistics were collected for six blocks of 1000 customers each. Unfortunately, this is not enough to produce very good accuracy, especially for the nodes with higher traffic intensities. (Compare with the simulation length in Section III.) The statistical reliability can be seen from the results of the six runs displayed in Fraker.[11] (These also appear in Appendix 1 of Whitt.[18]) An idea of the variability can also be seen from node 1 because all the approximations except the M/M/1 approximation are exact for node 1. When $\rho = 0.9$ the length of a 95-percent confidence interval approximately equals the estimated value; when $\rho = 0.7$ the length of

Table III—A comparison of approximations and simulation of the expected waiting time at each node in Fraker's model of eight queues in series: Case 1

Node No.	Traffic Intensity, ρ_j	Squared Coefficient of Variation, c_{aj}^2	Simulated Value	Approximation Methods				
				(Markovian Network) M/M/1	(Asymptotic Method) M/G/1	(Lag-1 Correlations) Fraker	QNA	
							EW_j	c_{aj}^2
1	0.7	1/8	0.98	1.63	0.92	0.92	0.92	1.00
2	0.5	1	0.30	0.50	0.50	0.38	0.38	0.61
3	0.5	0	0.19	0.50	0.25	0.13	0.16	0.71
4	0.7	1/4	0.73	1.63	1.02	0.62	0.63	0.58
5	0.3	0	0.01	0.13	0.07	0.01	0.01	0.42
6	0.9	1	7.50	8.10	8.10	6.03	5.56	0.40
7	0.9	1/8	3.91	8.10	4.55	4.16	4.09	0.89
8	0.3	1/4	0.00	0.13	0.08	0.01	0.01	0.33

Note: The arrival process is Poisson with rate 1 and the service-time distributions are Erlang.

Table IV—A comparison of approximations and simulation of the expected waiting time at each node in Fraker's model of eight queues in series: Case 2

Node No.	Traffic Intensity, ρ_j	Squared Coefficient of Variation, c_{aj}^2	Simulated Value	Approximation Methods (Markovian Network) M/M/1	(Asymptotic Method) M/G/1	(Lag-1 Correlations) Fraker	QNA EW_j	QNA c_{aj}^2
1	0.9	1	6.25	8.10	8.10	8.10	8.10	1.00
2	0.7	1/8	0.84	1.63	0.92	0.92	0.92	1.00
3	0.3	1/4	0.01	0.13	0.08	0.04	0.04	0.61
4	0.9	0	2.61	8.10	4.05	2.45	2.28	0.58
5	0.3	1/8	0.00	0.13	0.07	0.00	0.00	0.27
6	0.5	1/4	0.02	0.50	0.31	0.05	0.06	0.26
7	0.5	0	0.02	0.50	0.25	0.02	0.02	0.26
8	0.7	1	0.78	1.63	1.63	0.82	0.89	0.25

Note: The arrival process is Poisson with rate 1 and the service-time distrihutiona are Erlang.

a 95-percent confidence interval is about 25 percent of the estimated value.

Table V compares the approximations with simulation for the nodes with traffic intensity $\rho = 0.7$ in all eight cases. Since the approximations are exact for the first node, the first node is not included for the cases in which $\rho_1 = 0.7$ (Cases 1, 5, and 6). For the approximations, the difference between the approximation value and the simulation value is displayed.

Tables III through V show that QNA performs about the same as Fraker's approximation, which is based on lag-1 correlations and is especially designed for queues with Erlang service times. Both these approximations performed significantly better than the M/G/1 approximation, which in turn performs significantly better than the M/M/1 approximation.

Additional analysis of Fraker's models plus other queues in series is contained in Whitt.[18] The performance of QNA in these other cases is consistent with the description here.

V. A TIGHTLY COUPLED NETWORK OF TWO NODES

In this section we consider a two-node network analyzed by Kuehn[12] and Gelenbe and Mitrani.[13] This network is depicted in Fig. 2. It has one external arrival process, which comes to node 1. Customers completing service at node 1 leave the system with probability 1/2; otherwise they go to node 2 and then back to node 1 to be served again. At node 2 customers are immediately fed back to node 2 for another service with probability q_{22}, but in most cases $q_{22} = 0$.

We first consider Kuehn's experiment. There are eight cases with

Table V—The expected waiting time at the nodes with $\rho_j = 0.7$ in Fraker's eight cases of eight single-server queues in series

Case No.	Node No.	Simulated Value of EW_j	Approximation Methods			
			M/M/1	M/G/1	Fraker	QNA
1	4	0.73	+0.90	+0.29	−0.11	−0.10
2	2	0.84	+0.79	+0.08	+0.08	+0.08
	8	0.78	+0.85	+0.85	+0.04	+0.11
3	4	1.08	+0.55	+0.55	−0.04	+0.04
	8	0.55	+1.08	+0.37	0.00	−0.02
4	3	1.52	+0.11	+0.11	−0.04	−0.04
	6	0.02	+1.61	+0.80	+0.09	+0.16
5	3	0.74	+0.89	+0.28	+0.10	+0.11
6	5	0.33	+1.30	+0.49	+0.05	−0.01
7	4	0.78	+0.85	+0.14	−0.06	−0.02
	7	0.17	+1.46	+0.65	+0.02	−0.01
8	5	0.50	+1.13	+0.52	+0.04	+0.02
Average		0.67	+0.96	+0.43	+0.01	+0.03
Average Absolute Difference			0.96	0.43	0.05	0.06

Note: The value for the approximations is the approximate value minus the simulated value.

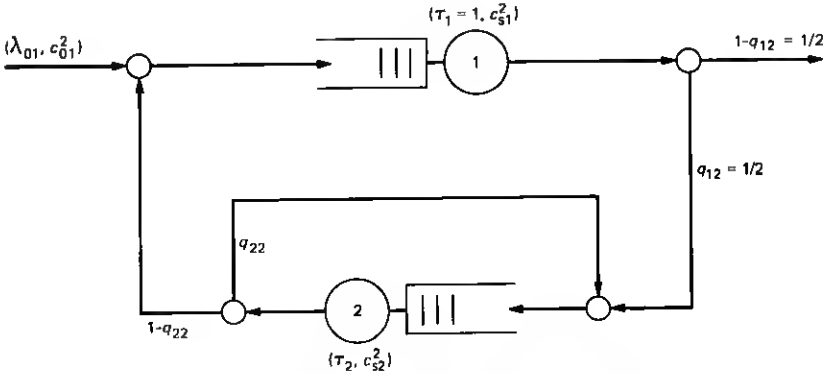

Fig. 2—Kuehn's first example: A network of two queues with one external arrival process.

three values of the external arrival rate for each case: $\lambda_{01} = 0.15$, 0.30, and 0.45. In each case the mean service time at node 1 is $\tau_1 = 1$ and the transition probability from node 1 to node 2 is $q_{12} = 1/2$, so that the traffic intensity at node 1 is $\rho_1 = 2\lambda_{01}$. For node 1 these three external arrival rates correspond roughly to light traffic ($\rho_1 = 0.3$), moderate traffic ($\rho_1 = 0.6$), and heavy traffic ($\rho_1 = 0.9$). In Cases 2 and 3 the traffic intensity at node 2 is the same as at node 1, i.e., $\rho_2 = 2\lambda_{01}$, but in all other cases it is $\rho_2 = \lambda_{01}$. In these other cases node 2 is always in relatively light traffic. The external arrival process is always a renewal process. Each interarrival-time distribution and service-time distribution is one of four distributions: deterministic (D with c^2

= 0), Erlang of order 4 (E_4 with $c^2 = 0.25$), exponential (M with $c^2 = 1$), or hyperexponential with balanced means (H_2^b witb $c^2 = 2.25$). The eight cases are indicated in Table VI. The system type is described by a triple such as $M/H_2/E_4$, which means tbat tbe interarrival-time distribution is M and the service-time distributions at nodes 1 and 2 are H_2 and E_4, respectively.

The results are described in Tables VII and VIII. The simulation results and Kuehn's approximation are taken from Kuehn.[12] Two different approximation results are given for QNA in Table VII. The first column is the standard application of QNA witb the network reconfigured to eliminate immediate feedback in the one case it occurs, at node 2 in Case 3. (See Section 3 of Ref. 1.) The final column of Table VII is an adjusted version of QNA to eliminate almost immediate feedback, which we discuss below.

For this network the quality of the standard QNA approximation is about the same as Kuehn's approximation. They both work well for low and moderate traffic intensities, e.g., about 10-percent average absolute relative percent error when $\rho_1 = 0.6$ (Table VIII), but not so well in heavy traffic. This two-node network presents an obvious difficulty for QNA. The network is tightly coupled so tbat many departures from node 1 rapidly return to node 1 for additional service. However, since these returning customers first pass through node 2, there is no immediate feedback, so that QNA does not reconfigure the network to eliminate the feedback. Nevertheless, it is evident that this almost immediate feedback for node 1 is very similar to immediate feedback and the potential exists for better results by reconfiguring the network to eliminate this feedback too.

In all cases except 2 and 3, almost immediate feedback is eliminated by applying the standard version of QNA with immediate feedback elimination twice. The first time we apply QNA to the full network and the second time we apply QNA with node 2 removed. Wben node

Table VI—The eight cases of Kuehn's two-node example in Fig. 2

System Number	System Type	Defining Parameters				
		c_{01}^2	τ_2	c_{s1}^2	c_{s2}^2	q_{22}
1	$M/H_2/E_4$	1.00	1.0	2.25	0.25	0.0
2	$M/H_2/E_4$	1.00	2.0	2.25	0.25	0.0
3	$M/H_2/E_4$	1.00	1.0	2.25	0.25	0.5
4	$M/H_2/H_2$	1.00	1.0	2.25	2.25	0.0
5	$M/E_4/E_4$	1.00	1.0	0.25	0.25	0.0
6	$M/D/D$	1.00	1.0	0.00	0.00	0.0
7	$H_2/H_2/E_4$	2.25	1.0	2.25	0.25	0.0
8	$E_4/H_2/E_4$	0.25	1.0	2.25	0.25	0.0

Notes: 1. In each case $\tau_1 = 1$ and $q_{12} = 1/2$.
2. In each case the arrival rate assumes one of three values: $\lambda_{01} = 0.15$, 0.30, and 0.45.

Table VII—A comparison of approximations and simulation of the expected total sojourn time (waiting time plus service time) in Kuehn's two-node network

System No.	External Arrival Rate, λ_{01}	Simulation (with 95-Percent Confidence Intervals)	Approximation Methods				
			M/M/1	M/G/1	Kuehn	QNA	QNA Adjusted
1	0.15	4.24 ± 0.08	4.04	4.50	4.51	4.51	4.24
	0.30	7.51 ± 0.29	6.43	8.13	8.22	8.26	7.25
	0.45	27.27 ± 5.63	21.82	32.67	33.47	34.10	27.35
2	0.15	5.95 ± 0.89	5.29	5.65	5.94	5.95	4.41
	0.30	10.91 ± 0.64	8.50	9.79	10.89	11.01	8.22
	0.45	49.49 ± 6.80	31.00	38.74	46.31	46.57	36.79
3	0.15	6.09 ± 0.19	5.29	5.65	6.11	6.12	4.48
	0.30	11.08 ± 0.63	8.50	9.79	11.60	11.66	8.25
	0.45	61.72 ± 17.99	31.00	38.74	51.27	51.43	35.07
4	0.15	4.50 ± 0.16	4.04	4.68	4.69	4.69	4.42
	0.30	7.96 ± 0.55	6.43	8.55	8.71	8.79	7.68
	0.45	29.91 ± 6.14	21.82	33.48	34.94	36.75	28.17
5	0.15	3.66 ± 0.05	4.04	3.63	3.63	3.64	3.81
	0.30	5.35 ± 0.16	6.43	5.13	4.93	5.00	5.67
	0.45	18.29 ± 3.27	21.82	14.67	12.80	12.80	18.12
6	0.15	3.43 ± 0.04	4.04	3.51	3.49	3.51	3.72
	0.30	4.83 ± 0.07	6.43	4.71	4.42	4.56	5.42
	0.45	13.59 ± 1.79	21.82	12.41	9.79	10.32	16.73
7	0.15	4.73 ± 0.12	4.04	4.50	4.60	4.62	4.78
	0.30	9.04 ± 0.56	6.43	8.13	8.59	8.92	9.14
	0.45	46.83 ± 10.64	21.82	32.67	35.88	39.74	39.50
8	0.15	3.67 ± 0.09	4.04	4.50	4.19	4.43	3.91
	0.30	5.78 ± 0.16	6.43	8.13	7.49	7.84	6.09
	0.45	17.46 ± 1.67	21.82	32.67	29.57	30.68	20.47

Notes: 1. In each case the traffic intensity at node 1 is $\rho_1 = 2\lambda_{01}$.
2. In Cases 2 and 3 the traffic intensity at node 2 is $\rho_2 = 2\lambda_{01}$; otherwise it is $\rho_2 = \lambda_{01}$.

2 is removed, the feedback to node 1 becomes immediate and the network is reconfigured by QNA to eliminate it. We use the second run with node 2 removed to determine the expected waiting time per visit at node 1. We use the first run to determine the expected number of visits to node 1 and the expected total sojourn time at node 2.

We do not treat nodes 1 and 2 symmetrically in Cases 1 and 4 through 8 because $\rho_1 = 2\rho_2$ so that ρ_2 is relatively small compared to ρ_1. Customers that return to node 1 via node 2 will not be delayed long at node 2 before coming back, but customers returning to node 2 via node 1 will be delayed relatively longer before coming back. If we had $\rho_1 < \rho_2$, we would remove node 1 in the second run of the QNA and focus instead on node 2.

In Cases 2 and 3 the traffic intensities at nodes 1 and 2 are equal, so the motivation for eliminating almost immediate feedback is less. What we have done for Table IV is first calculate the congestion

Table VIII—A comparison of approximation methods in Kuehn's two-node network with $\lambda_{01} = 0.3$ ($\rho_1 = 0.6$): The average absolute relative percent error in the expected total sojourn time compared with simulation

System Number	Approximation Methods					
	M/M/1	M/G/1	Kuehn	QNA Standard	QNA Adjusted	QNA Refined
1	−14.3	+8.3	+9.5	+10.0	−3.5	−3.5
2	−22.1	−10.3	−0.2	+0.9	−24.7	+0.9
3	−23.3	−11.6	+4.7	+5.2	−25.5	+5.2
4	−19.2	+7.4	+9.4	+10.4	−3.5	−3.5
5	+20.2	−4.1	−7.9	−6.5	+6.0	+6.0
6	+33.1	−2.5	−8.5	−5.6	+12.2	+12.2
7	−28.9	−10.1	−5.0	−1.3	+1.1	+1.1
8	+11.2	+40.7	+29.6	+35.6	+5.4	+5.4
Average Percent Error	21.5	11.9	9.4	9.4	10.2	4.7

measures for node 1 via the second run of QNA with node 2 removed as before. Then we use the results for node 1 to approximate the variability parameter of the arrival process to node 2. Finally, we analyze node 2 in isolation with the correct rates and this approximate arrival variability parameter. This works slightly better than the first procedure, but neither works well.

The results demonstrate that the standard version of QNA performs relatively well in Cases 2 and 3 when $\rho_1 = \rho_2$. The adjustment to eliminate almost immediate feedback yields a significant improvement when $\rho_1 > \rho_2$, but the results after adjustment to eliminate almost immediate feedback are much worse when $\rho_1 = \rho_2$.

As a refined procedure for this two-node network, we suggest eliminating almost immediate feedback at the node with higher traffic intensity when the traffic intensities differ significantly, and using the standard QNA algorithm otherwise. The refined procedure in Table VIII is standard QNA in Cases 2 and 3 and the adjusted QNA in all other cases.

Table VIII displays the relative percentage errors for all the approximations for the eight cases with $\lambda_{01} = 0.3$ ($\rho_1 = 0.6$). The refined procedure in the last column yields very good results. Table VIII also demonstrates that the standard version of QNA is significantly better than the M/M/1 approximation, but not uniformly better. In some cases, e.g., in Case 8, errors in opposite directions can cancel for the M/M/1 approximation.

The improvement from eliminating almost immediate feedback "by hand" suggests that it would be desirable to develop an automatic procedure for eliminating almost immediate feedback to incorporate

in QNA, and this is being investigated. It also indicates the potential for "tuning" QNA for particular applications.

We now consider Gelenbe and Mitrani's[13] experiment, which consists of five cases. The network is as depicted in Fig. 2 except the routing probability q_{12} is not exactly 1/2. The parameter values are given in Table IX and the results in Table X (pp. 137, 138 of Gelenbe and Mitrani[13]). We only include the best of three approximation schemes discussed by Gelenbe and Mitrani. Unfortunately, Gelenbe and Mitrani provided no information about the statistical reliability of the simulation estimates. Since ρ_1 and ρ_2 are nearly equal in each case, we did not try to eliminate almost immediate feedback.

The M/M/1 approximation values are obviously much too large because the M/M/1 approximation does not reflect the low variability of the service times. The M/G/1 approximation is much too low at node 2 because it does not benefit from the feedback elimination procedure. The Gelenbe-Pujolle procedure is better than the M/G/1 procedure, but not uniformly so. The QNA approximation is clearly

Table IX—The parameter values for Gelenbe and Mitrani's experiment with the two-node network in Fig. 2

Case No.	Parameter Values									
	λ_{01}	c_{01}^2	τ_1	c_{s1}^2	τ_2	c_{s2}^2	q_{11}	q_{12}	q_{21}	q_{22}
1	0.512	0.941	0.911	0.427	0.840	0	0	0.510	0.497	0.503
2	0.410	0.944	0.916	0.423	0.840	0	0	0.509	0.501	0.499
3	0.342	0.945	0.914	0.414	0.840	0	0	0.516	0.494	0.506
4	0.293	0.967	0.904	0.432	0.840	0	0	0.512	0.498	0.502
5	0.257	0.952	0.911	0.422	0.840	0	0	0.504	0.493	0.507

Table X—A comparison of approximations with simulations: The expected number of customers at each node in the two-node network in Gelenbe and Mitrani[13]

Case No.	Queue No.	Arrival Rate, λ_j	Traffic Intensity, ρ_j	Simulation Values	Approximation Methods			
					M/M/1	M/G/1	Gelenbe-Pujolle	QNA
1	1	1.04	0.952	13.82	19.83	14.14	11.96	11.98
	2	0.53	0.901	7.83	9.10	4.55	5.74	5.88
2	1	0.84	0.765	2.36	3.26	2.32	2.09	2.31
	2	0.43	0.713	1.87	2.48	1.24	1.69	1.84
3	1	0.71	0.646	1.60	1.82	1.29	1.20	1.41
	2	0.36	0.620	1.47	1.63	0.82	1.15	1.29
4	1	0.60	0.543	1.05	1.19	0.85	0.82	0.98
	2	0.31	0.519	1.01	1.08	0.54	0.78	0.90
5	1	0.52	0.472	0.76	0.89	0.63	0.62	0.76
	2	0.26	0.445	0.73	0.80	0.40	0.59	0.70

the best, but it may underestimate the congestion for high traffic intensities.

VI. KUEHN'S NINE-NODE NETWORK

We now consider a nine-node network analyzed by Kuehn,[12] which is depicted in Fig. 3. The mean service time at node j is $\tau_j = 1$ for each j. There are three external arrival processes with $\lambda_{0j} = 0.5$ for each j; these come to nodes 1, 2, and 3. Kuehn let tbe three external arrival processes be Poisson processes. As in Section V, all service-time distributions are D, E_k, M, or H_2^b. Kuehn considered two cases: homogeneous servers, in which all the service-time distributions are identical, and heterogeneous servers, in whicb nodes 1 tbrough 3 have one service-time distribution and nodes 4 through 9 have anotber.

Kuehn compared his approximation with the M/M/1 and M/G/1 approximations and simulation for service-time variability parameters ranging from $c_{sj}^2 = 0$ to 4. He focused on the expected total sojourn time (waiting time plus service time) in tbe network and the expected sojourn time per visit in node 4. For this network Kuebn found, first, that the service-time variability parameters are significant (the sojourn-time measures increase significantly with c_{sj}^2); second, that his approximation tracks the simulation well; and, third, that the M/G/1 approximation also works well, but not quite as well as his approximation.

We obtained similar results applying QNA. The QNA approximation values are indistinguishable from tbe approximation values displayed graphically by Kuehn, which are consistently within the sim-

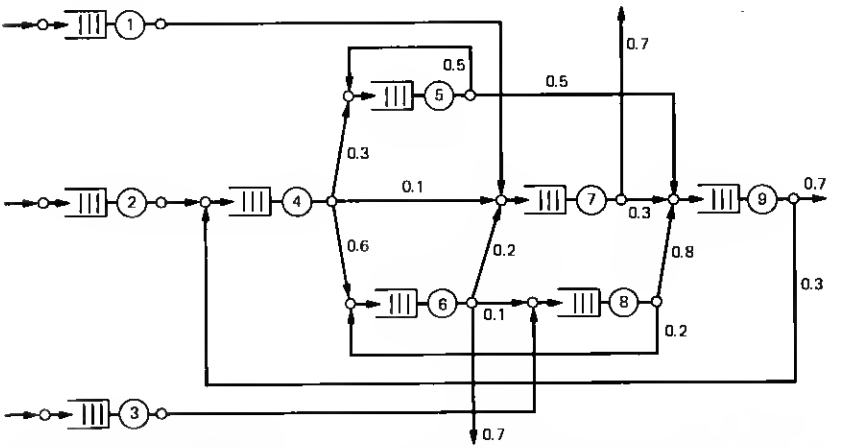

Fig. 3—Kuehn's second example: A network of nine queues with three Poisson external arrival processes.

ulation confidence intervals. In Tahle XI we display some of our results. Here we consider only tbe case of homogeneous servers in which $c_{sj}^2 = 0$, 1, or 4. However, we let the variahility parameter of all external arrival processes by $c_{0j}^2 = 0.25$, 1.0, or 4.0. We thus obtain nine cases.

Kuehn did simulations only in tbe case $c_{0j}^2 = 1$. The different values ohtained hy QNA when $c_{0j}^2 \neq 1$ suggest that the congestion measures in this model are more sensitive to the variability of the service times than the variahility of the interarrival times. Of course, we should expect that the M/G/1 approximation might perform well when the variability parameters of the external arrival processes are 1 or close to 1, but for relatively large and interconnected networks the M/G/1 approximation may perform well for other external arrival processes. It performs reasonably well here when $c_{0j}^2 = 0.25$ or 4.0.

VII. A COMPUTER SYSTEM MODEL: INPUT BY ROUTES

In this section we apply QNA to a network with input by customer classes and routes as in Section 2.3 of Ref. 1. We compare QNA to a simulation model used in the development of a computer system at Bell Lahoratories. In this model there are five nodes and two customer classes.

The customer classes correspond to typical functions performed by the system. The route for each class represents a typical sequence of operations performed by the system to process one of these functions.

Table XI—Approximations of expected sojourn times in Kuehn's nine-node network in Fig. 3: The case of homogeneous servers, c_{sj}^2 identical for all j

Expected Total Sojourn Time in the Network		Service-Time Variability Parameters		
		$c_{sj}^2 = 0$	$c_{sj}^2 = 1$	$c_{sj}^2 = 4$
QNA	$c_{0j}^2 = 0.25$	6.1	11.5	28.3
	$c_{0j}^2 = 1.00$	7.5	12.9	29.6
	$c_{0j}^2 = 4.00$	12.3	17.7	34.4
M/G/1		8.2	12.9	27.1
M/M/1		12.9	12.9	12.9

Expected Sojourn Time per Visit in Node 4		Service-Time Variability Parameters		
		$c_{sj}^2 = 0$	$c_{sj}^2 = 1$	$c_{sj}^2 = 4$
QNA	$c_{0j}^2 = 0.25$	1.76	3.71	9.64
	$c_{0j}^2 = 1.00$	2.34	4.25	10.15
	$c_{0j}^2 = 4.00$	4.38	6.29	12.19
M/G/1		2.62	4.25	9.12
M/M/1		4.25	4.25	4.25

In the simulation model, both the routes and the service times at the nodes are deterministic for each class. Hence, the input for QNA is just as specified in Section 2.3 of Ref. 1 with the service-time variability parameters set equal to 0, i.e., $c^2_{skj} = 0$ for each class k and j on class k's route. The two routes we consider are given in Tables XII and

Table XII—The data for the first route in the model of Section VII

Number	Node	Mean Service Time	Number	Node	Mean Service Time
1	2	4	46	4	30
2	1	10	47	1	16
3	1	66	48	4	30
4	1	60	49	1	36
5	1	106	50	1	38
6	1	65	51	1	45
7	4	30	52	4	30
8	1	16	53	1	16
9	4	30	54	4	30
10	1	36	55	1	36
11	1	12	56	1	15.5
12	1	40	57	1	45
13	4	30	58	4	30
14	1	16	59	1	16
15	4	30	60	4	30
16	1	36	61	1	16
17	1	62	62	4	30
18	1	63	63	1	16
19	1	42	64	4	30
20	1	14.5	65	1	36
21	3	550	66	1	38
22	5	0.01	67	1	45
23	3	50	68	4	30
24	1	10	69	1	16
25	1	8	70	4	30
26	1	20.5	71	1	36
27	1	55.5	72	1	27
28	1	42	73	1	25
29	1	14.5	74	1	40
30	3	550	75	4	30
31	5	0.01	76	1	16
32	3	50	77	4	30
33	1	10	78	1	36
34	1	8	79	1	13
35	1	20.5	80	1	58
36	1	63.5	81	1	8
37	1	76	82	2	16
38	4	30	83	5	0.01
39	1	36	84	2	4
40	1	15.5	85	1	10
41	1	45	86	1	36
42	4	30			
43	1	16			
44	4	30			
45	1	16			

XIII. Note that node 1 frequently appears several times in succession, so that there is immediate feedback at node 1. Also note that the service times differ at different visits to the same node.

The customer classes arrive according to independent Poisson processes. In the case we consider the arrival rates of Classes 1 and 2 are 0.00015278 and 0.00030555, respectively.

The QNA, M/G/1, and M/M/1 approximations are compared with simulation in Table XIV. The simulation values are the average of three separate runs. The values from these separate runs are displayed to give an idea of the statistical reliability. The congestion measures compared are the expected waiting times at the nodes and the expected total waiting time (excluding service time) on three route segments. The first segment is the first 25 nodes of the second route; the second segment is the first 21 nodes on tbe first route; and the third segment is eight nodes from node 23 to node 30 on the first route. In Table XIV the waiting times at the nodes are measured in milliseconds while the waiting times on the route segments are measured in seconds.

From Table XIV, it is apparent that QNA with immediate-feedback elimination performs reasonably well, significantly better than the M/M/1 and M/G/1 approximations. Since the approximating variability parameters of the arrival processes are very close to 1, QNA without immediate-feedback elimination is very similar to the M/G/1 approximation. Hence, again we see that eliminating immediate feed-

Table XIII—The data for the second route in the model of Section VII

Number	Node	Mean Service Time	Number	Node	Mean Service Time	Number	Node	Mean Service Time
1	2	4	21	4	30	41	1	46
2	1	10	22	1	36	42	1	14.5
3	1	66	23	1	42	43	3	400
4	1	60	24	1	14.5	44	1	8
5	1	106	25	3	400	45	1	12
6	1	65	26	5	0.01	46	1	30
7	4	30	27	3	350	47	1	58
8	1	16	28	1	10	48	1	8
9	4	30	29	1	16	49	2	16
10	1	36	30	1	20.5	50	5	0.01
11	1	12	31	1	84	51	2	4
12	1	65	32	1	45	52	1	10
13	4	30	33	4	30	53	1	36
14	1	16	34	1	16			
15	4	30	35	4	30			
16	1	36	36	1	16			
17	1	62	37	4	30			
18	1	40	38	1	16			
19	4	30	39	4	30			
20	1	16	40	1	36			

Table XIV—A comparison of approximations and simulation of the expected waiting times for the model of Section VII

Method		Expected Waiting Time at the Nodes				Total Expected Waiting Time on a Route Segment		
		Node 1	Node 2	Node 3	Node 4	1 (25 nodes)	2 (21 nodes)	3 (8 nodes)
Simulation runs	1	51.9	0.058	224.1	2.11	1.25	1.12	0.66
	2	58.5	0.058	245.6	2.27	1.37	1.27	0.77
	3	57.6	0.055	236.2	2.17	1.35	1.23	0.74
Simulation average		56.0	0.057	235.3	2.18	1.32	1.21	0.72
M/M/1		59.2 (+5.7)	0.09	402.5 (+71.1)	6.53 (+200.0)	1.45 (+9.8)	1.32 (+9.1)	1.16 (+61.1)
M/G/1		43.5 (−22.3)	0.07	245.5 (+4.2)	3.26 (+49.5)	1.01 (−23.5)	0.91 (−24.8)	0.75 (+4.2)
QNA (eliminating feedback)		50.2 (−10.3)	0.07	244.1 (+3.7)	2.81 (+28.9)	1.11 (−15.9)	1.01 (−16.5)	0.79 (+9.7)
		Additional Information About the Network						
		1	2	3	4	1	2	3
Mean service time		33.1	8.0	350.0	30.0	1.28	1.31	0.75
Traffic intensity		0.64	0.11	0.54	0.18	---	---	---
c^2 of the service time from QNA		0.47	0.50	0.22	0.00	---	---	---
c^2 of the arrival process from QNA		0.92	1.00	0.99	0.90	---	---	---

Notes: 1. The relative percent errors appear below the approximation in parentheses.
2. The value $c_{a1}^2 = 0.47$ at node 1 is before adjustment for feedback; after adjustment it is 0.79.
3. The units of measurement are milliseconds for the nodes and seconds for the route segments.

back helps. The M/M/1 approximations at nodes 3 and 4 evidently are too large because the service times are nearly constant. However, at node 1 the M/M/1 approximation does pretty well, apparently because two different errors cancel. (We have not displayed the relative percentage errors at node 2 since it seems of little consequence because of the low traffic intensity.)

It is interesting to know how the system would perform if the external arrival processes are not Poisson and if the service times at the nodes on the routes are not deterministic. With QNA we can easily perform such sensitivity analyses. We can simply change the variability parameters of the external arrival processes and the service times on the routes. The results of such a study are given in Tables XV and XVI. Table XV gives the approximating variability parameter of the

arrival process at each node as a function of the external arrival process c^2 and the service time c^2. It is assumed that botb external arrival processes have the same c^2 and that all service times on hoth routes have the same c^2. From Table XV, it is evident that the variability of the external arrival process hardly has any effect. Tbus, QNA predicts that this model, and perhaps the system itself, will be robust to changes in the variability of the arriving traffic. The variability from outside is evidently dissipated on tbe long routes through the network.

Tahle XVI descrihes the impact of changing the service-time variahility at the nodes on the routes. The service-time variability at the nodes would increase significantly and, thus, QNA predicts that the expected waiting times would also increase significantly. Simulations to test these predictions are planned.

Table XV—The approximate variability parameter of the arrival process at each node determined by the QNA, as a function of the given variability parameters: The model of Section VII

External Arrival Process c^2	Node	Service Time c^2 at Each Node on the Route				
		0.0	0.2	0.5	1.0	2.0
$c^2 = 1$	1	0.9155	0.9506	1.0034	1.0912	1.2669
	2	0.9995	0.9997	1.0000	1.0001	1.0016
	3	0.9937	0.9966	1.0010	1.0083	1.0230
	4	0.8950	0.9480	1.0274	1.1598	1.4245
$c^2 = 2$	1	0.9162	0.9513	1.0040	1.0918	1.2676
	2	1.0086	1.0088	1.0092	1.0097	1.0107
	3	0.9938	0.9967	1.0011	1.0084	1.0231
	4	0.8953	0.9483	1.0277	1.1601	1.4249
$c^2 = 4$	1	0.9175	0.9527	1.0053	1.0932	1.2689
	2	1.0268	1.0271	1.0274	1.0279	1.0290
	3	0.9939	0.9968	1.0012	1.0085	1.0231
	4	0.8959	0.9488	1.0283	1.1607	1.4254

Table XVI—The approximate service-time variability parameter c_{sj}^2 and mean delay EW_j at node j determined by the QNA, as a function of the variability of each service time on the route: The model of Section VII

Node Characteristic	Node	Service Time c^2 at Each Node on the Route				
		0.0	0.2	0.5	1.0	2.0
c_{sj}^2	1	0.787	0.905	1.082	1.377	1.968
	2	0.500	0.800	1.250	2.000	3.500
	3	0.220	0.464	0.830	1.441	2.661
	4	0.000	0.200	0.500	1.000	2.000
EW_j	1	50.2	54.9	61.6	72.6	94.6
	2	0.07	0.08	0.10	0.13	0.20
	3	244.1	293.3	367.7	491.4	739.0
	4	2.81	3.72	4.95	6.87	10.78

VIII. A PACKET-SWITCHED COMMUNICATION-NETWORK MODEL

In this section we consider a model of a packet-switched communication network analyzed in Section 4.3.1 of Gelenbe and Mitrani.[13] The basic model has 5 switching nodes and 12 one-way data links, as depicted in Fig. 4. However, in this model each data link is a server and the packets waiting for transmission on the link form the queue. Packets are assumed to arrive at the switching nodes according to independent Poisson processes. Each packet arriving from outside at node i has final destination j with probability d_{ij}. Each packet with destination j goes next to node r_{ij} from node i. Hence, there is a fixed route for each origin-destination pair.

We analyze this network using the input by classes and routes in Section 2.3 of Ref. 1. However, unlike Section VII, here the service-time parameters are associated with the nodes rather than the routes (which is an input option in QNA). The network of queues has 12 nodes with one server at each node and 20 routes. As specified by Gelenbe and Mitrani, the service rate at nodes 1, 2, 7, 8, 11, and 12 is 4.8 (in thousands of bits per second) and the service rate at the other nodes is 48. Since packet lengths are assumed constant, $c_{ij}^2 = 0$ for all j.

For this example the matrices $D \equiv (d_{ij})$ of destination probabilities and $R = (r_{ij})$ of next-node routes are:

$$D = \begin{bmatrix} 0.00 & 0.10 & 0.20 & 0.10 & 0.60 \\ 0.40 & 0.00 & 0.40 & 0.15 & 0.05 \\ 0.10 & 0.20 & 0.00 & 0.60 & 0.10 \\ 0.30 & 0.30 & 0.30 & 0.00 & 0.10 \\ 0.10 & 0.25 & 0.30 & 0.35 & 0.00 \end{bmatrix} \quad R = \begin{bmatrix} 0 & 3 & 3 & 3 & 2 \\ 4 & 0 & 5 & 5 & 4 \\ 6 & 6 & 0 & 9 & 8 \\ 10 & 10 & 10 & 0 & 12 \\ 1 & 1 & 7 & 11 & 0 \end{bmatrix}.$$

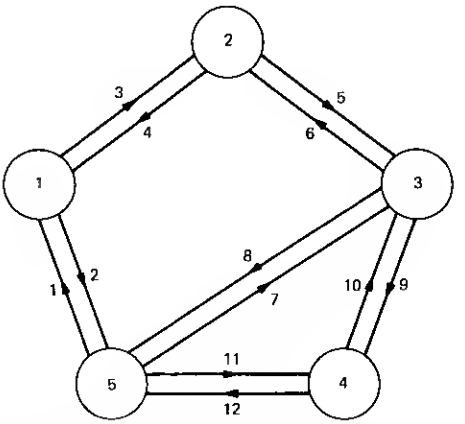

Fig. 4—Gelenbe and Mitrani's model of a packet-switched network: The 12 links are the nodes in the network of queues.

Tbe input data for the routes are given in Table XVII. Tbese data are obtained from the matrices D and R plus the external arrival rates of 6.00, 8.25, 7.50, 6.75, and 1.50 at the five switching nodes (p. 141 of Gelenbe and Mitrani[13]).

The results are compared with Gelenbe and Mitrani's approximation and simulation in Table XVIII. Since only 6000 packets reached their destination in the simulation, the statistical reliability of tbe simulation estimates cannot be very good, cf. Section III. Our analysis is revealing. First, Gelenbe and Mitrani describe their results as average buffer queue lengths, which might be thought to exclude the customer (packet) being served (transmitted). However, the results obviously include the customer in service. Second, the two M/M/1 approximations should agree, but they do not. Evidently, the arrival rates at switches 1 and 2 were not actually as cited in the text.[13] Hence, tbe numbers for the one heavily loaded link, link 2, cannot be meaningfully compared.

It is useful to consider the expected number waiting excluding the customer in service. This is easily obtained because the probability that the server is busy is exactly the traffic intensity, ρ (see Section 11.3 of Heyman and Sobel[14]). We thus obtain estimates of the expected number waiting by subtracting ρ from the numbers displayed in Table

Table XVII—The input data by routes for Gelenbe and Mitrani's model of a packet-switched communication network

Route Number	Origin-Destination Pair	External Arrival Process Parameters $\hat{\lambda}_k$	c_k^2	Number of Nodes on the Route	Node Sequence
1	1,2	0.60	1	1	3
2	1,3	1.20	1	2	3,5
3	1,4	0.60	1	3	3,5,9
4	1,5	3.60	1	1	2
5	2,1	3.30	1	1	4
6	2,3	3.30	1	1	5
7	2,4	1.24	1	2	5,9
8	2,5	0.41	1	2	4,2
9	3,1	0.75	1	2	6,4
10	3,2	1.50	1	1	6
11	3,4	4.50	1	1	9
12	3,5	0.75	1	1	8
13	4,1	2.03	1	3	10,6,4
14	4,2	2.03	1	2	10,6
15	4,3	2.03	1	1	10
16	4,5	0.68	1	1	12
17	5,1	0.15	1	1	1
18	5,2	0.38	1	2	1,3
19	5,3	0.45	1	1	7
20	5,4	0.53	1	1	11

XVIII. When this is done, some of the simulation estimates become negative, demonstrating that the input parameters are incorrect or the statistical reliability of the simulation is not very good. In this case, probably both problems exist.

We also observe that the traffic intensities at all link queues but the second are very small, so the numbers displayed in Table XVIII are mostly estimates of the traffic intensities themselves. Moreover, because the traffic intensities are small, the variability parameter of the departure process produced by QNA will be very close to the variability parameter of the arrival process (see Section 4.5 of Ref. 1). This would not be true for the second link with traffic intensity 0.835, but note that all departures from the second link leave the system. Since the external arrival processes are all Poisson, QNA should and does perform virtually the same as an M/G/1 approximation. In fact, since the service times are all constant ($c_{sj}^2 = 0$), the approximation reduces to the M/D/1 system. Moreover, we predict that a proper simulation of this model with the specified parameters will yield values very close to the M/D/1 approximation.

We also display in Table XIX the point-to-point (origin-destination) average total service times, delays, and sojourn times (service times plus delays) produced by QNA. No simulation values were available for comparison, however. The output is useful to indicate unacceptably high or low values. It is also useful to determine the separate contributions of service times and delays to sojourn times. Of course, in

Table XVIII—A comparison of approximations and simulation: The expected number waiting and being served on each link in the Gelenbe-and-Mitrani model of a packet-switched communication network depicted in Fig. 4

			Approximation Methods			
			M/M/1			
Link No.	Traffic Intensity, ρ_j	Simulation Value	Gelenbe and Mitrani	via QNA	Gelenbe-Pujolle	QNA
1	0.110	0.117	0.123	0.124	0.116	0.117
2	0.835	1.920	3.000	5.076	1.875	2.955
3	0.058	0.132	0.139	0.061	0.131	0.060
4	0.135	0.163	0.170	0.156	0.157	0.146
5	0.132	0.105	0.127	0.152	0.125	0.142
6	0.131	0.173	0.157	0.151	0.146	0.141
7	0.094	0.087	0.104	0.103	0.099	0.099
8	0.156	0.208	0.185	0.185	0.171	0.171
9	0.132	0.155	0.162	0.152	0.147	0.142
10	0.127	0.129	0.145	0.145	0.136	0.136
11	0.110	0.106	0.123	0.124	0.116	0.117
12	0.142	0.154	0.164	0.165	0.152	0.153

Table XIX—Average point-to-point service times, delays, and sojourn times for the Gelenbe-and-Mitrani model of a packet-switched communication network

Route No.	Mean Total Service Time on Route	Poisson Arrivals $c^2 = 1.0$		Bursty Arrivals $c^2 = 4.0$	
		Mean Total Delay on Route	Mean Total Sojourn Time on Route	Mean Total Delay on Route	Mean Total Sojourn Time on Route
1	0.021	0.001	0.021	0.001	0.022
2	0.042	0.002	0.044	0.003	0.044
3	0.063	0.004	0.066	0.005	0.068
4	0.208*	0.529*	0.737*	1.920*	2.128*
5	0.021	0.002	0.022	0.002	0.023
6	0.021	0.002	0.022	0.002	0.023
7	0.042	0.003	0.045	0.005	0.046
8	0.229*	0.530*	0.759*	1.922*	2.151*
9	0.042	0.003	0.045	0.004	0.046
10	0.021	0.002	0.022	0.002	0.046
11	0.021	0.002	0.022	0.003	0.024
12	0.208*	0.019	0.228*	0.077	0.285*
13	0.063	0.005	0.067	0.006	0.068
14	0.042	0.003	0.045	0.004	0.046
15	0.021	0.002	0.022	0.002	0.023
16	0.208*	0.017	0.226*	0.069	0.277*
17	0.208*	0.013	0.221*	0.025	0.233*
18	0.229*	0.014	0.243*	0.026	0.255*
19	0.208*	0.011	0.219*	0.043	0.251*
20	0.208*	0.013	0.221*	0.052	0.260*

Note: The larger values are marked with an asterisk.

Table XIX delays play a significant role only for routes using the second link.

We conclude by remarking that the assumption of Poisson arrivals for packets at each switch made by Gelenbe and Mitrani[13] often is not realistic. Often messages containing many packets arrive according to a Poisson process, but the packets arrive in a much more bursty manner. Hence, it is appropriate to use QNA with arrival-process variability parameters much larger than 1. The last two columns of Table XIX give the mean delays and sojourn times when the variability parameters of the external arrival processes are changed from $c^2 = 1.0$ to $c^2 = 4.0$. When this is done here, the large delays on routes 4 and 8 increase significantly. To a large extent, QNA was motivated by the need to be able to systematically study the effect of such variability.

IX. ACKNOWLEDGMENTS

I am grateful to Anne Seery for using QNA to generate much of the data here, as well as for writing the QNA program. I am grateful to E. B. Zucker for the simulation data in Section VII.

REFERENCES

1. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., this issue.
2. W. Whitt, "On Approximations for Queues, I: Extremal Distributions," B.S.T.J., *63*, No. 1, Part 1 (January 1984).
3. J. G. Klincewicz and W. Whitt, "On Approximations for Queues, II: Shape Constraints," B.S.T.J., *63*, No. 1, Part 1 (January 1984).
4. W. Whitt, "On Approximations for Queues, III: Mixtures of Exponential Distributions," B.S.T.J., *63*, No. 1, Part 1 (January 1984).
5. W. Whitt, "The Marshall/Marshall and Stoyan Bounds for IMRL/G/1 Queues are Tight," Oper. Res. Letters, *1*, No. 6 (December 1982), pp. 209–13.
6. W. Whitt, "Refining Diffusion Approximations for Queues," Oper. Res. Letters, *1*, No. 5 (November 1982), pp. 165–9.
7. W. Kraemer and M. Langenbach-Belz, "Approximate Formulae for the Delay in the Queueing System GI/G/1," *Congressbook*, Eighth International Teletraffic Congress, Melbourne, Australia, 1976, pp. 235, 1–8.
8. S. L. Albin, *Approximating Queues with Superposition Arrival Processes*, Ph.D dissertation, Department of Industrial Engineering and Operations Research, Columbia University, 1981.
9. S. L. Albin, "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues," Department of Industrial Engineering Rutgers University, 1982.
10. S. L. Albin, "On Poisson Approximations for Superposition Arrival Processes in Queues," Management Sci., *28*, No. 2 (February 1982), pp. 126–37.
11. J. R. Fraker, *Approximate Techniques for the Analysis of Tandem Queueing Systems*, Ph.D. dissertation, Department of Industrial Engineering, Clemson University, 1971.
12. P. J. Kuehn, "Approximate Analysis of General Queueing Networks by Decomposition," IEEE Trans. Commun., *COM-27*, No. 1 (January 1979), pp. 113–26.
13. E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems*, New York: Academic Press, 1980.
14. D. P. Heyman and M. J. Sobel, *Stochastic Models in Operotions Research*, Volume I, New York: McGraw-Hill, 1982.
15. W. Whitt, "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," Oper. Res., *30*, No. 1 (January–February 1982), pp. 125–47.
16. G. Marsaglia, K. Ananthanarayanan, and N. Paul, "Random Number Generator Package–Super Duper," School of Computer Science, McGill University, 1973.
17. W. Whitt, "Queues with Superposition Arrival Processes in Heavy Traffic," unpublished work, 1982.
18. W. Whitt, "Approximations for Departure Processes and Queues in Series," Navy Res. Log Qtrly., to be published.

AUTHOR

Ward Whitt, A. B. (Mathematics), 1964, Dartmoutb College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At Bell Laboratories he is in tbe Operations Research Department in the Network Analysis Center.